Pers Ubiquit Comput (2008) 12:223–235
DOI 10.1007/s00779-007-0148-1

ORIGINAL PAPER

Evaluating meeting support tools

Wilfried M. Post · Mirjam A. A. Huis in 't Veld · Sylvia A. A. van den Boogaard

Received: 17 February 2006 / Accepted: 30 October 2006 / Published online: 6 March 2007

Abstract Many attempts are underway for developing meeting support tools, but less attention is paid to the evaluation of meetingware. This article describes the development and testing of an instrument for evaluating meeting tools. First, we specified the object of evaluation—meetings—by means of a set of input, process, and outcome factors. Then, we designed the process of evaluation, consisting of, first, the generation of meeting behavior in the form of a controlled series of meetings within the context of a project and, second, the measurement of the identified meeting factors. To measure these factors, a rating scale, questionnaires, and information flow analysis were used. Next, the instrument was tested, and the factors for successful meetings were determined in 13 projects in which four participants had to meet four times. The evaluation instrument proved to be a reliable and useful aid for the development and improvement of meeting tools.

Keywords Evaluation · Meetings · Design · Teams · Meetingware

1 Introduction

Professionals spend 20–40% of their time at work in meetings [1]. Meetings are essential to structure and coordinate work in organizations; the costs considering time, money and logistics are however staggering and satisfaction levels are low [2]. Often the obtained result is hardly worth the costs and effort. Meetings are meant to be sources of stimulation, support, and solutions; they should fulfill organizational as well as personal needs. For some of us, though, meetings are an annoyance and a waste of time. For all its difficulty, teamwork is still essential and technology is often considered as the Holy Grail to improve the meeting process and outcomes [3]. Can this situation indeed be improved using technology?

The research reported here is carried out as part of AMI, a European Integrated Project that aims at developing new multimodal technologies that support human interaction in the context of face-to-face and remote meetings (see http://www.amiproject.org). These technologies enable storage, interpretation, and retrieval of captured meeting interactions. They are the basis for the development of a multimodal meeting browser, which should make meetings more effective, efficient, and pleasurable. However, to know how such a browser *may* improve meetings, we need to explore the determinants for successful meeting behavior. And to know whether our browser, or any other meeting tool, indeed *does* improve meetings, we need an evaluation instrument. For both purposes, we have to make meetings measurable. In this article, we describe the development and testing of this evaluation instrument and how we use it for exploring the determinants of successful group meetings.

All evaluations have common features: in all cases there is an *object* being evaluated, a *process* through which one or more attributes are judged and valued, and all evaluations have a *purpose* [4]. The purpose of *our* evaluation instrument is to estimate the quality of meetings and provide insight into the factors that contribute to this quality level, in order to direct improvements.

W. M. Post (✉) · M. A. A. Huis in 't Veld · S. A. A. van den Boogaard
TNO Human Factors, Kampweg 5, 3769 DE Soesterberg, The Netherlands
e-mail: Wilfried.Post@tno.nl

M. A. A. Huis in 't Veld
e-mail: Mirjam.HuisintVeld@tno.nl

With this study, we contribute to the large research track on computer-based collaboration. A specific category of computer-based meeting support developed and studied in this area is often referred to as group support systems (GSS) [3, 5–7], but the term electronic meeting systems is also applied. In line with [2], we will use meetingware to refer to this type of systems. Meetingware can be considered a subcategory of groupware: software that supports collaboration in groups. Both for groupware as well as for meetingware, several evaluation instruments are developed. A number of these evaluation instruments focus on dispersed groups [8, 9]. We focus on groups in collocated settings. Some evaluation approaches are developed in the tradition of software evaluation, whereas other approaches are the result of research into group dynamics. Here we focus on group related factors and not so much on the evaluation of technology related factors. The evaluation instrument we aim to develop should enable measuring group factors and especially the influence that group processes have on meeting outcomes. As a result, the evaluation instrument should enable us to measure the effect of supporting technology on these processes and outcomes. So far, few studies have focused on the interplay among process variables and meeting outcomes [5].

The object and process of our evaluation instrument are specified in the next two sections. In the subsequent two sections, the resulting evaluation instrument is the object of evaluation itself (Sect. 4 and 5). We investigate the reliability, validity, usefulness, and acceptability of the evaluation instrument by testing it on a large set of meetings. With the resulting data, we also explore the determinants of successful group meetings. We conclude this article discussing the merits and drawbacks of our evaluation instrument.

2 Object of evaluation

The object of evaluation concerns all aspects that make up a meeting; our evaluation instrument should support the assessment of these aspects. They are specified in this section. Studying meeting behavior means studying work groups. Many theories on work group functioning use the conceptual input-process-outcome model as a frame of reference. It relates a set of contextual determinants, which affect work group effectiveness through the mediation of internal activities of work groups [10–12]. Process measurements are often referred to as measures of performance, and outcome measures are often referred to as measures of effectiveness. Outcome measures assess the quantity and quality of the end result [13]. These can be distinguished from process measures that describe the strategies, steps, or procedures used to accomplish a task.

Brodbeck [10] poses that performance is an aggregate of those behaviors that are relevant for achieving the goals specified. Examples of these kinds of behaviors are effort, supportiveness, and team performance functions. Effectiveness is the degree to which the performance outcomes approach the goals specified. Productivity is how efficiently a particular level of effectiveness is achieved. A better understanding of the relationship between performance and effectiveness—that is, a better understanding of predictor–criterion relationships—can contribute to making meetings more effective, for optimizing teamwork in meetings involves using processes that on average will lead to more favorable outcomes.

In evaluating meetings, we consider meetings not as isolated events (in line with [14]). A meeting is usually preceded and followed by individual work such as preparation for the next meeting, distribution of the results, and execution of actions that have been agreed upon. Then, a next meeting is held, etc. Together, the meetings contribute to the achievement of a higher goal, such as a project goal or a mission objective. In our view, meetings should therefore be measured with respect to this ''higher goal.'' Figure 1 illustrates this process. Figure 2 gives an overview of the variables we use in this study. These variables are discussed next.

2.1 Input variables

Post, Cremers and Blanson Henkemans [14] distinguish as input variables means, methods, individual factors, team factors, task factors, organizational factors, and environmental factors. Means refer to systems and tools that support a meeting-related task (e.g., an interactive large screen display). Methods refer to prescriptions of how to do a particular task (e.g., a procedure to chair a meeting). A team is a group of individuals who see themselves and are seen by others as a social entity [15], which is also the case for the participants of a meeting (e.g., a management team). Team processes are influenced by individual characteristics [16], in particular the different roles that the

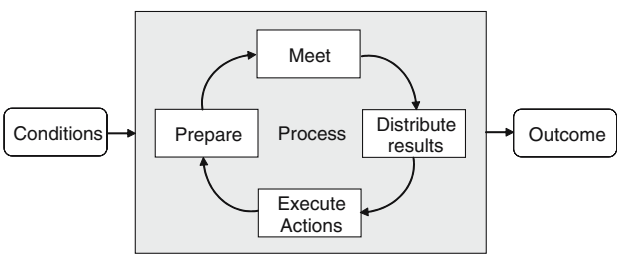

Fig. 1 Meetings as part of a cyclic process with a higher goal

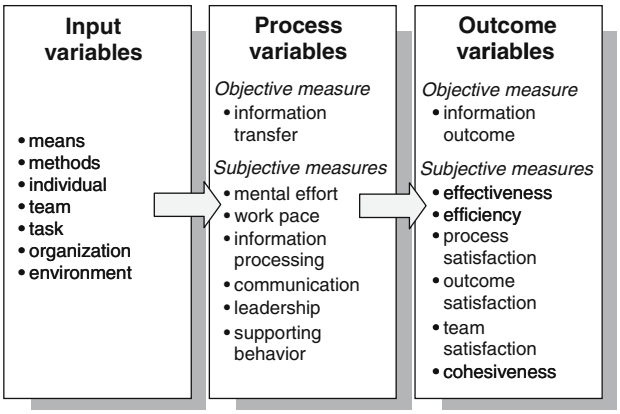

Fig. 2 A framework for studying meeting behavior

individuals play [17] (e.g., the chair). The task refers to the work that must be done to reach certain goals. Through the task, team members become interdependent [18]. Tasks can be described as individual tasks [19] (e.g., design) or as group tasks [20] (e.g., negotiate). Organizational factors refer to aspects, such as organizational structure and culture [21]. Environmental factors refer to aspects external to the organization, such as the market. The success of teams and organizations strongly depends on how they manage the unexpected dynamics of the environment [22].

2.2 Process variables

The amount of *information transfer* is an important process variable. Limited information transfer can be the consequence of strong solution directedness and a quick opinion formation [23]. Groups tend to come up with solutions at an early stage, although the problem is hardly clear. An analysis of the nature of the problem and of the factors that support the problem will not take place, nor will an analysis of the pros and cons of the solutions that are brought in. This is caused by an opinion formation of the group that is too quick, by which some solutions are immediately rejected and vanish from the process of weighting. Limited information transfer can also be the consequence of conformity and self-censorship. Because of the consequences of the amount of information transfer on the outcome of a meeting, we take this variable into account in our research.

We are also interested in the *mental effort* needed during the meetings. It is easy to imagine that the better the performance, the more effort it takes. It is also known that the amount of mental effort needed increases during the day because people get tired. Finally, it is important to care for the mental well being of the participants of a meeting. Participants should not do too much work, nor too less to let them feel well (see e.g., [24]). A related aspect we want to take a look at is *work pace*. We wonder to what extent the participants of a meeting feel that they have to work under time pressure. Having to work too hard may affect the outcome of the task.

In literature, several process variables are mentioned that are crucial to effective teamwork. Smith-Jentsch et al. [13], for example, describe an approach to develop and evaluate a human-performance measurement system designed to assess training needs called team dimensional training (TDT). TDT is based on years of team research, it describes, diagnoses, and evaluates processes that lead to effective outcomes. TDT helps to obtain a better insight into the process of teamwork. At the end of an exercise, the participants evaluate their own team performance under guidance of a facilitator, with the aim of improving their performance. The evaluation takes place in the light of four dimensions, which are thoroughly checked on reliability and validity. The dimensions are subdivided into concrete and observable components. The four dimensions are information exchange, communication, supporting behavior, and leadership. Information exchange is the extent to which every member of the team knows what information should be transferred and to whom and when. For our purpose, we extend this variable to include other aspects of processing information: whether this information is remembered and used. We call this extended variable *information processing*. The *communication* dimension refers to the way in which information is passed through (e.g., understandability). *Supporting behavior* is the mechanism by which team members compensate for one another's weaknesses by correcting errors and shifting workload. *Leadership* is concerned with those behavioral characteristics that direct the team in a certain direction. Each member of the team can guide the team by taking the lead. The four dimensions are taken because they are reliable and valid, and represent superior teamwork strategies.

2.3 Outcome variables

The first outcome variable we consider as important is *information outcome*. Where *information transfer* refers to the amount of information individuals share in the meeting, *information outcome* refers to information that is not only shared but also actually used in reaching a desirable outcome. The second outcome variable we take into account is *team effectiveness*. Whether groups are more effective than individuals depends on the criteria that are used for defining effectiveness [25]. Group decisions tend to be more accurate (better-quality decisions), whereas individuals are faster. Groups tend to be more creative and group decisions tend to score better on acceptance of the final solution. The choices groups make are more creative and accurate, because groups bring more complete information and knowledge to a decision, so they generate more ideas.

In addition, the give-and-take that typically takes place in group decision processes provides diversity of opinion and increases the likelihood that weak alternatives will be identified and abandoned.

According to Robbins [25], effectiveness cannot be considered without also assessing *efficiency*. Individual decision makers are almost always more efficient than groups. Group decision making consumes more work hours. In deciding whether to use groups, it should be considered whether increases in effectiveness are larger than the losses in efficiency.

An important goal that must be achieved in every meeting is that the participants of the meeting are content with the process and outcomes of the meeting. Robbins [25] states that organizations with more satisfied employees tend to be more effective than organizations with less satisfied employees. A consistent negative relationship is found between satisfaction and absenteeism, and between satisfaction and turnover. Moreover, researchers with strong humanistic values argue that satisfaction is a legitimate objective of an organization. Organizations should be concerned not only with the quantity of work life—that is, concerns such as higher productivity and material acquisitions—but also with its quality. Organizations have a responsibility to provide employees with jobs that are challenging and intrinsically rewarding. Therefore, satisfaction is considered as important. We distinguish satisfaction with the process, the outcome, and the team. *Process satisfaction* refers to the satisfaction with decisions made and how they were made. O*utcome satisfaction* refers to the quality of the outcome. For a design project, for instance, this is the completeness, price, correctness, and precision of the final solution. The last kind of satisfaction we distinguish is *team satisfaction*, which concerns satisfaction with the team process and the achievement as a team.

Cohesiveness refers to the degree of interpersonal attraction and liking among team members. To assess cohesiveness, researchers almost always ask team members to indicate their personal feelings about other members and/or their liking of the team as a whole [26]. Positive affect promotes helping behavior and generosity, cooperation, and a problem solving orientation during negotiations. When positive affect occurs in the form of attraction to team members, it may translate into greater motivation to contribute fully and perform well as a means of gaining approval and recognition. Positive affect is likely to be particularly beneficial for improving performance. This is especially the case in problem-solving situations, where flexible and creative thinking can lead to more effective resolutions than compromise. One way to enhance group cohesiveness is increasing the time members spend together [24].

3 Process of evaluation

In this section, we describe the measurement instruments we propose to address all aspects of the framework. First, we will present the process measures and subsequently the outcome measures. In both categories, a distinction is made between objective and subjective measures. Next, we explain the evaluation procedure.

3.1 Process measures

3.1.1 Objective measure

We propose to measure *information transfer* by counting the number of items of information that individual participants have available and share in the meeting. In this particular research, we only considered information shared by PowerPoint presentations. The total number of items of rightly transferred information was divided by the total number of transferable items of information and multiplied by 100 to obtain the percentage of rightly transferred information per team.

3.1.2 Subjective measures

Mental effort can be measured with the rating scale mental effort (RSME) [27]. The RSME is a translation of the Dutch *Beoordelings Schaal Mentale Inspanning* (BSMI). We adopt the translation of the BSMI in [28]. The score is indicated by digits on the left side of a scale from 0 to 150, the variable anchors are indicated by words on the right side. The anchors range from ‘‘absolutely no effort’’ to ‘‘extreme effort.’’

For the aspects *information processing* and *leadership*, we have developed four items for each variable, which can be added to questionnaires presented at the end of a meeting and presented at the end of the unit of meetings. For the variables c*ommunication* and *supporting behavior*, we have developed four items that can be added to a questionnaire about a whole unit of meetings, such as a whole project or a whole mission.

3.2 Outcome measures

3.2.1 Objective measure

We propose to measure the objective outcome measure *information outcome* by comparing all the information that the individuals have available and is required to share to reach a desirable outcome with what the participants actually share and use. In this particular research, we used the percentage of correctly applied design requirements as the information outcome measure. As will become clear

below, the participants will meet in the context of a design project. At the end of the project, the participants have to evaluate their result (i.e., a design prototype) according to the design requirements. These requirements have to be gathered and exchanged during the project.

3.2.2 Subjective measures

For measuring *effectiveness*, we have constructed items that request opinions on the quality of the solution in the light of team versus individual work. For *efficiency*, the items asked the subjects whether they think the job could have been done faster with fewer people. For *cohesiveness*, the items asked the participants how much they liked the other members. Items on how hard and fast participants had to work have been constructed to measure *work pace*. Further, we constructed items for measuring *process*, *outcome*, and *team satisfaction*.

3.3 Evaluation procedure

Together, the measurement instruments constitute the evaluation instrument we propose. The evaluation procedure is as follows. The objective measurements are done by the researchers afterwards. The subjective measures are collected by the participants themselves by means of rating scales and questionnaires. The subjective measurements are not all taken after every meeting, because this may influence the workload and meeting behavior of the participants. For example, questions about supporting behavior after each meeting may stimulate this behavior in the following meeting. Only those variables that are expected to change over time are inquired after every meeting. Table 1 provides an overview of the evaluation instrument.

The items are scored on seven-point rating scales. The extremes on the seven-point rating scales were ''not applicable at all'' at the left end, and ''very much applicable'' at the right end. (An overview of all items is provided in [29].)

4 Testing the instrument

For testing our evaluation instrument, we need to collect a large set of comparable meeting data. For generating meeting behavior in a natural yet controlled and replicable manner, we will make use of an experimental paradigm, developed by Post et al. [14]. In this paradigm, a team of four participants, who act as employees of a consumer electronics company, has to design a prototype for a new TV remote control in four alternating individual and meeting sessions. The setting and scenario used to test the evaluation instrument have several elements that contribute to a realistic setting. The group members are individually

Table 1 Evaluation instrument

When	What	How
Analyse afterwards (about all meetings)	Information transfer	Percentage of rightly shared information
	Information outcome	Percentage of correctly applied design requirements
Collect before each meeting	Mental effort	150 pt scale
Collect after each meeting	Mental effort	150 pt scale
	Info processing	4 items
	Leadership	4 items
	Process satisfact.	3 items
	Cohesiveness	5 items
Collect at the end (about all meetings)	Info processing	4 items
	Leadership	4 items
	Process satisfact.	3 items
	Cohesiveness	5 items
	Workpace	4 items
	Communication	4 items
	Support. behavior	8 items
	Effectiveness	4 items
	Efficiency	7 items
	Outcome satisfact.	5 items
	Team satisfaction	2 items

instructed and trained for their group role. In addition, information necessary for the design is distributed among them. The scenario further has elements of a dynamic market (changing fashion) and an organizational context (a particular company, budget costs). New information about the market and the organization is sent to individual participants. The marketing expert, for example, receives information about changes in the market and related changes in the evaluation criteria for the design. By simulating the dynamics of these contexts the team functions in a less isolated setting (see [14] for more details). We will use this design project for measuring the process and outcomes of meetings.

4.1 Research questions

Our research question is formulated as follows: Is the evaluation approach we propose a useful and reliable instrument for measuring meetings in terms of process and outcome? This question is divided into three specific questions:

1. Is the evaluation instrument useful and reliable for measuring process changes?
2. Is the evaluation instrument useful and reliable for measuring outcome changes?
3. Is the evaluation instrument useful and reliable for predicting the outcome of the process?

5 Participants

Of the 52 participants, 46 were undergraduate students of information science at the University of Utrecht. These students took a course in groupware. Participation in our research was compulsory to obtain course credits. Most of these subjects were more or less acquainted before the start of the project. The remaining six participants were undergraduate students of various other studies and participated voluntarily. All participants were paid € 60 for about 7 h of work. The age of the subjects ranged from 18 to 28 years, the mean age was 21 years. Of the subjects, 46 were male and 6 female. Four participants made up one design team. So in total, 13 design teams were formed. Participants did not have prior knowledge of their roles or tasks.

5.1 Apparatus

To simulate an office environment, we made use of four private office rooms and one meeting room. In the meeting room, two large interactive screens were set up next to each other. One was used as an electronic white board and the other one for PowerPoint presentations.

The participants had to work with a laptop computer, which they ported from their office to the meeting room and vice versa. On their laptops, they could use common office tools: MS Word, MS PowerPoint and MS Excel, the email application Outlook Express, and the web browser Internet Explorer (to look at simulated web pages). The laptops were wireless connected. Documents produced could be saved in a private folder and/or in a shared folder on the desktop.

A separate room served as an observatory for the researchers. The participants were observed and recorded by means of video cameras in each room and wireless head-mounted microphones. Their laptop interactions and the interaction with the large screen displays were monitored via RealVNC. The large screen displays were automatically captured in real time, after any significant change. A server laptop was used to run a scenario application, especially developed for this purpose. This application controlled the scenario automatically by sending emails and alerts, or opening web pages, for a specific role at specified points in time. Email was used to send information about the design problem as well as instructions and questionnaires. The server laptop also collected the data of the digital questionnaires and all documents that the participants produced.

5.2 Material

The design team consisted of a project manager, an industrial designer, a user interface designer, and a marketing expert. They were explained that the design project has four phases: project kick-off, functional design, conceptual design, and detailed design. In the functional design phase it is determined what needs and desires of users are to be fulfilled, what effect the apparatus should have, and how the apparatus works to fulfill its functions. In the conceptual design phase, the conceptual specification of components, properties and materials, and the conceptual specification of the user interface are determined. At the same time, trend watching should take place. In the detailed design phase, the look-and-feel of the prototype must be determined and evaluated. In each design phase, each team member has a specific task. For a detailed description of the precise content of all information provided to the subjects and an overview of the whole scenario data base, we refer to [30].

In each phase, the participants had to work first individually and then together. When working individually, they received role-specific, detailed project specifications and instructions, sufficient to prepare the next meeting. This information consisted of descriptions of the task, inspirations by presenting other similar designs, reports of market research, pre-structured PowerPoint slides, and the like.

5.3 Procedure

Figure 3 shows a time schedule of the scenario. For pragmatic reasons, each design project was carried out in one day.

Having arrived at the TNO Human Factors lab in the morning, the participants received oral instructions and a handout on how to use the tools. The participants were told to speak and write in English because of the international use of the data. The scenario started at about 10:00 a.m. As can be seen in Fig. 3, the first individual session (that took about 30 min) was followed by a first meeting (30 min). This was followed by a second individual session (30 min) and a second meeting (45 min). At 12:15 p.m., there was a lunch break of about 45 min. They were instructed not to talk about the research. After lunch, there was a third individual session (30 min) and a third meeting session (45 min). The project concluded with a fourth individual (30 min) and meeting (45 min) session. In the fourth individual session, the industrial designer and the user interface designer worked together.

In the individual work sessions, the participants read their email, visited simulated web pages, executed new tasks, and prepared their PowerPoint presentation on their findings. The project manager had to document the former meeting. During meeting sessions, the team members showed the PowerPoint presentations to each other, discussed the findings, and made decisions on various aspects of the new TV remote control prototype. The scenario application controlled by alerting pop-ups when to go to the meeting room and to start the meeting, and when to end the meeting and to go to the private workspace again.

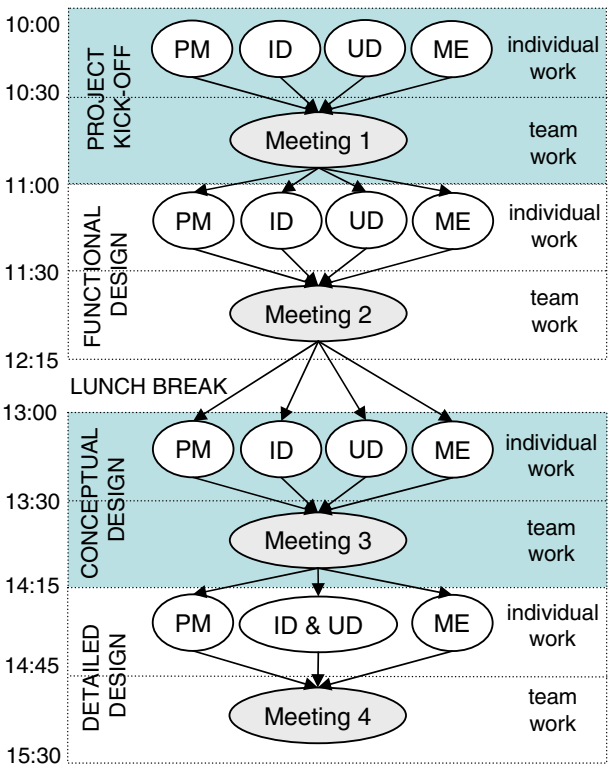

Fig. 3 Meeting cycle: Time schedule for individual and meeting activities and measurements. *PM* project manager; *ID* industrial designer; *UD* user interface designer; *ME* marketing expert

5.4 Applying the evaluation instrument

All proposed instruments that together constitute the evaluation instrument were applied in the test. After every individual work session and after every meeting session, the participants were asked by email to fill out a *mental effort* rating scale. After every meeting session, they were also asked to fill out a questionnaire about the meeting, on the process measures, *information processing* and *leadership*, and the outcome measures, *process satisfaction* and *cohesiveness*. After the final meeting, the project manager was asked to report on the design product. All participants were also asked to fill out a questionnaire about the whole project, on the process measures, *work pace*, *communication, and supporting behavior*, and on the outcome measures, *effectiveness, efficiency*, *outcome satisfaction*, and *team satisfaction*.

For the objective process measure *information transfer*, we count the number of items that the individual participants are provided and the number they use in their PowerPoint presentations to the group. The total number of items of transferred information divided by the total number of transferable items of information multiplied by 100 represents the percentage of rightly transferred information per team.

For the objective outcome measure *information outcome*, the amount of rightly formulated criteria at the end of the project is taken as an outcome measure. We will compare the list of requirements that are given during the day to individual team members, with the list of requirements that the team comes up with. The percentages correct criteria are taken as a measure.

6 Results

In this section, we present the results of the data analyses, organized around the research questions presented in Sect. 4.1.

6.1 Reliability testing

To test the reliability of our questionnaires, we computed Cronbach's alpha. *Information processing* shows low alphas between 0.36 and 0.45. Therefore, we will not use this

scale in further analyses. *Leadership* in meeting 1 ($\alpha = 0.59$) and *communication* ($\alpha = 0.58$) are moderately reliable. We think that this is still sufficient to consider these variables for further analyses. Looking at the Cronbach's alphas, if an item was deleted from a scale, it did not yield a substantial increase in any alpha of any scale, so the original items were maintained. In sum, we can conclude that the scales of most questionnaires are reliable measures to gauge meeting behavior, except for the scale measuring *information processing*.

6.2 Descriptive statistics

Descriptive statistics of the various variables were examined. All the variables that are scored on a seven-point scale have a mean score somewhere between 4 and 6. Apart from *workspace* ($M = 4.10$), the statements about *efficiency* ($M = 4.78$) are least applicable to the teams compared to the other variables. *Cohesiveness* statements are most applicable ($M = 6.01$). *Outcome satisfaction* ($M = 4.86$) scores are lower than *process satisfaction* ($M = 5.84$) and *team satisfaction* ($M = 5.76$). Mean *process satisfaction* increases with every meeting and scores highest at the end of the day. Mean *cohesiveness* decreases with every meeting, increases in meeting 4, but is highest at the end of the day as well. *Leadership* first increases, then decreases, and then again increases: it is highest at the end of the day. *Mental effort* in the first individual session ($M = 39.71$) is lower than in the next individual sessions ($M = 66.15, 56.25, 61.83$, respectively). The same applies to *mental effort* in the meeting sessions ($M = 42.69$ vs. 62.21, 63.52, 65.50, respectively). In other words, this means that the first individual session took the participants ''some'' effort and the next sessions ''rather much'' to ''considerable'' effort. The same applies to the meeting sessions. The differences in effort between the individual and the meeting sessions do not seem big. In total, both kind of sessions took ''rather much'' effort. Whether the means described above differ significantly will be discussed later in this section. On average, 58% of the information was transferred by the teams. On average, 61% of the right evaluation criteria were used at the end of the designing project to evaluate the final design.

6.3 Usefulness

6.3.1 Can we measure process changes?

Let us first investigate changes in the process variable *mental effort.* In particular, we are interested in whether we can measure an increase of *mental effort* due to fatigue Two different *mental effort* indications are distinguished: one for the individual work sessions and one for the meeting sessions. Line charts that represent the mean scores on these variables at the respective moments of measurement are given in Fig. 4. To answer the question whether the scores of the variables on the first moment of measurement differ from the last moment, we perform a non-parametric Wilcoxon signed ranks test, instead. (A paired sampled t test cannot be performed, because the variables are not normally distributed.) The mean score on the RSME after the first individual session ($M = 39.71$, ''some effort'') is significantly lower than the mean score on the RSME after the last individual session ($M = 61.83$, ''rather much effort''), $Z = -4.59$, $p < 0.001$. The mean score on the RSME after the first meeting session ($M = 42.69$, ''some effort'') is significantly lower than the mean score on the RSME after the last meeting ($M = 65.50$, ''considerable effort'') as well, $Z = -4.92$, $p < 0.001$.

To test the differences of the means of the variables between the succeeding moments of measurements, a single group univariate repeated measures analysis is conducted. For *mental effort*, the tests of within-subjects effects are significant for both the individual and the meeting sessions, $F(2.61, 133.14) = 17.38$, $p < 0.001$ and $F(2.88, 146.77) = 21.05$, $p < 0.001$, respectively. In the analysis, we requested repeated contrasts as planned comparisons, which compare session 1 with session 2, session 2 with session 3, and session 3 with session 4. The tests of within-subjects contrasts reveals that for the individual sessions *mental effort* in the first session ($M = 39.71$) is lower than in the second session ($M = 66.15$), $F(1, 51) = 65.72$, $p < 0.001$, and that in the second session ($M = 66.15$) is higher than that in the third ($M = 56.25$), $F(1, 51) = 7.00$, $p < 0.05$. The third and the fourth sessions do not differ significantly. For the meeting sessions it applies that just the *mental effort* in the first session ($M = 42.59$) differs significantly from the second session ($M = 62.21$), $F(1, 51) = 30.33$, $p < 0.001$.

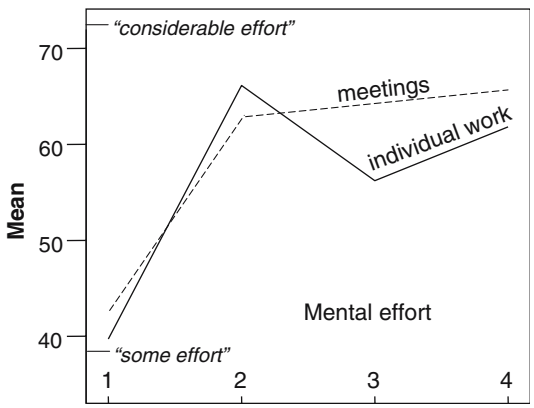

Fig. 4 Line charts of the mean scores on *mental effort* at individual and meeting session 1, 2, 3, and 4

Another question is whether we can measure differences between sessions. To evaluate the differences in means between the mean score on the RSME of the individual sessions and that of the meeting sessions, a paired-samples t test was used, since both variables are normally distributed. The result of the t test revealed no significant difference.

Still another question is if we can measure whether the different role players experienced the same amount of mental effort. To test whether it took the project manager, the industrial designer, the user interface designer, and the marketing expert on average the same amount of effort to fulfill their tasks, a one-way analysis of variance with Tukey's tests as post hoc tests, was performed with the *mental effort* variable per session as the dependent variable and the project role as factor.

For the individual session, it applies that there is just a significant difference in the second session. The mean of the project managers ($M = 57.31$, ''rather much effort'') and the mean of the user interface designers ($M = 55.00$, ''rather much effort'') are lower than the mean of the industrial designers ($M = 81.92$, ''great effort''), $F(3, 48) = 3.90$, $p < 0.05$. For the meeting sessions no significant differences (at $\alpha = 0.01$) were found.

Can we also measure that the better the process, the more *mental effort* it takes? To test whether a higher score on the process variables cohere with a higher score on *mental effort* in the meeting sessions, a Spearman bivariate correlation analysis is done. The only significant correlation that is found is that between *mental effort* and *work pace*, $\rho = 0.25$, $p < 0.05$.

So, we are indeed able to measure differences and changes in mental effort, although we did not find those in all cases we looked at.

A second process variable we would like to be able to measure the changes in is *leadership*. The line charts that represent the mean scores on the variable *leadership* at the various moments of measurement are given in Fig. 5, together with that of the outcome variables s*atisfaction* and *cohesiveness*, discussed in the next subsection. To answer the question whether the scores on the variables in the first meeting differ from those in the last meeting, a Wilcoxon signed ranks test is executed. *Leadership* in the first meeting ($M = 4.99$) is significantly lower than in the last meeting ($M = 5.60$), $Z = -3.90$, $p < 0.001$.

To test the differences of the means of the variable between the succeeding moments of measurements, again, a single group univariate repeated measures analysis is conducted. The tests of within-subjects effects reveal that there is a difference for *leadership*, $F(2.09, 106.78) = 8.30$, $p < 0.001$.

In the analysis, we requested repeated contrasts as planned comparisons, which compare meeting 1 with meeting 2, meeting 2 with meeting 3, and meeting 3 with meeting 4. The tests of within-subjects contrasts show that *leadership* in the first meeting ($M = 4.99$) is significantly lower than in the second meeting ($M = 5.35$), $F(1, 51) = 6.13$, $p < 0.05$, and that *leadership* in the third meeting ($M = 5.25$) is significantly lower than in the fourth meeting ($M = 5.60$), $F(1, 51) = 15.91$, $p < 0.001$.

So, also for *leadership* we have been able to find changes.

6.3.2 Can we measure outcome changes?

Is our instrument capable, for example, of determining whether the more time participants spend together, the higher will be the scores on outcome variables such as *satisfaction* and *cohesiveness*? Line charts that represent the mean scores on the variables *satisfaction* and *cohesiveness* at the various moments of measurement are given in Fig. 5. To know whether the scores on the variables in the first meeting differ from the last meeting, a Wilcoxon signed ranks test is executed. *Satisfaction* in the first meeting ($M = 5.25$) is significantly lower than in the last meeting ($M = 5.68$), $Z = -2.58$, $p < 0.05$. The results did not show any difference in *cohesiveness*, $Z = 0.63$, $p = 0.53$.

To test the differences of the means of the variables between the succeeding moments of measurements, again a single group univariate repeated measures analysis is conducted. The tests of within-subjects effects reveal that there is a difference in *satisfaction*, $F(2.78, 141.65) = 2.96$, $p < 0.05$, but not in *cohesiveness*, $F(2.75, 144.09) = 1.70$, $p = 0.17$.

In the analysis, we requested repeated contrasts as planned comparisons, which compare meeting 1 with meeting 2, meeting 2 with meeting 3, and meeting 3 with meeting 4. The tests of within-subjects contrasts show that

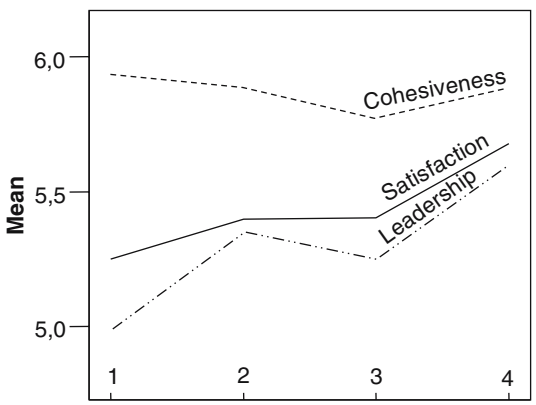

Fig. 5 Line charts of the mean scores on process variable *leadership* and outcome variables *satisfaction* and *cohesiveness* at meeting session 1, 2, 3, and 4

satisfaction in the third meeting ($M = 5.40$) is significantly lower than in the fourth meeting ($M = 5.68$), $F(1, 51) = 4.15$, $p < 0.05$.

Summarizing, we showed that we can indeed measure changes in satisfaction, but we did not show this for cohesiveness.

6.3.3 Can we predict the outcome from the process?

A multiple regression analysis was performed with the outcome variables as dependent variables, and the process variables that significantly correlate with the outcome variables as independent variables.

A multivariate outlier was found to be an influential point in the regressions with *outcome satisfaction* as dependent variable. The logbook showed in this case that the project manager behaved in an extremely dominant way, which had not been the case in the other teams. We removed this point from the regression analysis in which *outcome satisfaction* is the dependent variable, making N for this regression 51 instead of 52.

Table 2 shows the results of the regression analyses (method stepwise). The p-values have been halved since the directions of the regression coefficients are predicted. For *effectiveness*, *supporting behavior* accounts for 18% of the variance. For *efficiency*, *leadership* accounts for 23% of the variance. *Process satisfaction* can be predicted from the variables *leadership*, *communication,* and *supporting behavior*. These three variables account for 51% of the variance. *Leadership* contributes most to the regression ($\beta = 0.31$), followed by *communication* ($\beta = 0.28$) and *supporting behavior* ($\beta = 0.26$). *Outcome satisfaction* can be predicted only from *supporting behavior* (27% variance accounted for). *Team satisfaction* can be predicted from *leadership and supporting behavior*, accounting for 52% of the variance. *Leadership* ($\beta = 0.41$) contributes somewhat more than *supporting behavior* ($\beta = 0.37$). *Cohesiveness* can be predicted from *leadership* and *supporting behavior* as well, accounting for 70% of the variance. Again, *leadership* ($\beta = 0.47$) contributes somewhat more than *supporting behavior* ($\beta = 0.43$). The process variable w*ork pace* does not contribute to any of the outcome variables.

In order to compute correlations for the variables *information transfer* and *information outcome*, all variables were aggregated to team scores ($N = 13$). The process variable *information transfer* does not correlate significantly with any of the outcome variables. The outcome variable *information outcome* correlates significantly with the process variables *leadership*, $\rho = 0.52$, $p < 0.05$; *communication*, $\rho = 0.66$, $p < 0.01$; and *supporting behavior*, $\rho = 0.50$, $p < 0.05$.

Finally, we can conclude also that with our instrument, we are able to predict the outcome from the process.

7 Conclusions and discussion

The goal of this study was to develop an instrument for evaluating meeting support tools. We defined the scope of evaluation not as a single meeting but as a cycle of meetings, such as a project. Next, we based this instrument on an input-process-outcome model, and specified the model in a large number of factors. For almost all factors, we constructed measurement instruments. The resulting evaluation instrument consisted of an already existing rating scale, one new questionnaire with 16 items, another one with 50 items, and an information flow analysis instruction. Together, the instrument measured 19 factors. Next, we tested the evaluation instrument in 13 natural, yet controlled, meeting cycles, each involving four meetings. We found that leadership, supporting behavior and communication are important predictors for successful meetings. We therefore conclude that these factors need to be taken into account in future research on meeting behavior.

However, generalizations from this student-based study to the workplace has its limitations. Although the setting

Table 2 Multiple regression analyses with outcome variables as dependant variables and process variables as predictors ($N = 52$)

Outcome variables	F	df, df res.	R^2	C	Process variables (information processing removed)					
					Leadership		Communication		Supporting behavior	
					B	β	B	β	B	β
Effectiveness	11.27**	1, 50	0.18	2.04	–	–	–	–	0.53**	0.43**
Efficiency	14.49***	1, 50	0.23	1.77	0.53***	0.47***	–	–	–	–
Process satisfaction	16.31***	3, 48	0.51	1.39	0.29*	0.31*	0.25*	0.28*	0.27*	0.26*
Outcome satisfaction ($N = 51$)	18.27***	1, 49	0.27	0.04	–	–	–	–	0.87***	0.52***
Team satisfaction	26.37***	2, 49	0.52	0.79	0.44**	0.41**	–	–	0.44**	0.37**
Cohesiveness	55.83***	2, 49	0.70	1.33	0.42***	0.47***	–	–	0.42***	0.43***

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

and scenario applied have several elements that contribute to a realistic setting, the effect of some factors could not be tested or balanced for in the design. A first relevant limitation is the gender composition of the teams: almost all teams consisted of males only. Traditionally, engineering has been a majority male field; however, as more and more females enter the field, mixed gender teams are likely to become more common [31]. In some studies, it was assumed that the combination of different skills that males and females bring to the team task results in a higher level of performance. However, a test of this premise has resulted in contradicting outcomes [31–33]. As a consequence, the effect of the team's gender composition on our outcomes cannot easily be predicted. In general, however, women have a more interactive, people-oriented and cooperative work style, whereas men have more analytical decision-making tendencies and competitive orientation [32, 33] and this may affect the scores on communication and leadership.

Another limitation concerns the team members' lack of experience with the design task. In a real-world setting, it is likely that design experts also contribute to a design assignment and several studies show that senior designers apply different strategies in a design task (see e.g., [34]). It is further likely that larger differences in seniority affect scores on leadership.

A third limitation is the focus in this study on design teams, whereas eventually the evaluation instrument should be suitable for evaluating all types of teams and meetingware supporting these teams. Although this focus functions as a strength as well, as it enables comparison between teams, it limits the generalization of our outcomes.

To overcome the limitations of this study, the next step would be a test of the evaluation instrument in a real-world setting, e.g., real (design) teams working on an assignment. It would be interesting to see whether our conclusions can be validated in these settings. Are leadership, supporting behavior, and communication indeed important predictors for successful meetings? However, as this study consists of many realistic elements and overcomes the general tendency to focus on single meetings in an isolated context [7], we confidently anticipate such a test.

We will now discuss the results of the evaluation instrument test in terms of the following evaluation criteria: reliability, validity, acceptability, and usefulness. The limitations described above also hold for outcomes related to these aspects of the evaluation instrument.

7.1 Reliablity and validity

An instrument is reliable if it produces consistent results. We have checked the reliability of the questionnaires constructed in this study. Except for *information processing*, they are all sufficiently reliable. However, further research may be devoted to increase the reliability scores and improve the individual measurement instruments. The information flow analysis was carried out in this study by only one observer. So, a second observer is needed to test the inter-observer reliability of the variables *information transfer* and *information outcome*.

A question that one may ask is whether meetings by students represent real meetings. In our opinion, they do in our set-up. The students acted as real participants with their own individual roles and worked together well to reach their common goal. Furthermore, the fact that we could measure the presence of factors such as leadership, supporting behavior, and cohesiveness contributes to the conclusion that we had to do with real teams with common team processes.

A second question that one may ask is whether the instrument as proposed here is complete. Does it cover all possible factors? We can only conclude that further research is needed to state whether other processes contribute to successful outcomes as well, and whether other outcome variables can be identified and measured. We expect that, certainly in the context of a design project, economic measures could be identified. Ideally, we would economically value the products of each team: their developed remote control prototypes. We could, for example, ask domain experts to assess each design solution. More work remains to be done in that area.

A general remark is that we should realize that for a final validity test of our individual instruments, we need an objective standard, which is, unfortunately, not available.

7.2 Acceptability and usefulness

The evaluation instrument, as proposed by us, proves to be useful in providing insight into factors that determine the quality of meetings. In addition, the instrument is useful to assess to what extent factors determine this quality. However, applying the instrument takes much time and effort, especially for the evaluator(s). To generate reliable measurements, ten teams of four team members are needed in each condition. Although it is possible to automatically generate data by letting respondents fill out the questionnaire on-line, this still implies a lot of effort. Moreover, information flow analysis requires a lot of effort for the researchers. On the other hand, an alternative is not readily available.

The evaluation instrument we developed can be used to compare different meeting conditions by manipulating the various input variables. Here we mention some of the many possible manipulations.

Means: First of all, with the availability of the AMI multimodal meeting browser, we are now able to compare

alternative designs or evaluate the browser by comparing meetings with and without the browser. Another manipulation could be that, e.g., when the project manager is situated at a distance, he can communicate with his team by email or video conferencing, which has become more common in present-day companies. In this example, especially the influence on leadership would be interesting to examine.

Method: Leadership is found to be an important factor, but what type of instructions or training are the best (e.g., authoritarian vs. democratic)? In our research, we instructed the project manager to pay explicit attention to team building. The effect of such instructions, or instruction to the other members to take initiative when needed, or any other training in leadership can be measured with the AMI meeting browser. Another manipulation is the distribution of available time over the preparatory phase and the meeting phase. The question then is how preparation contributes to an effective, efficient, and satisfactory meeting.

Individual and team: In our data collection, a selection took place on a particular individual characteristic: we asked each team to assign a project manager to itself. Most of the time, a person with the best leading capacities took up that role spontaneously, or was recognized as such and pushed by the others. To find out whether everybody can become a leader, this assignment can take place differently. Another aspect is the size of the team, which is known to influence team performance [35]. Could certain roles be combined, e.g., the user interface designer and the industrial designer, without loss of performance?

Task: This research was centered on a design task. For other tasks, such as negotiation, planning, and crisis management, other results may be found.

Organization and environment: In this category, an interesting manipulation is underway. At two different sites, the University of Edinburgh, UK, and IDIAP, Martigny, Switzerland, the same data are collected. We are now in a process of comparing the meetings, and are curious whether there are cultural differences, differences in workload due to native speech, among other issues.

In the past decades, there has been a growing tendency to work in groups [36]. Organizations and organizational research have a continually changing nature. Because of technological, social, and other developments, teams in organizations in the near future may be quite different from teams as they are now. These changes will mount continuing challenges to group researchers and theorists. We think that an important contribution of our work is the insight that attention should be given to *the factors of success* in the future design of meetingware. To start with, the multimodal meeting browser we are developing in the AMI project, how can good leadership, supporting behavior, and communication be maintained using a meeting tool, irrespective of the circumstances and specific use of the tool?

Acknowledgments This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multiparty Interaction, FP6-506811, publication AMI-238). Anita Cremers and three anonymous reviewers provided valuable comments on an earlier version of this article.

References

1. Napier RW, Gershenfeld MK (1993) Groups: theory and experience. Houghton Mifflin Company, Boston
2. Antunes P, Costa CJ (2003) Perceived value: a low cost approach to evaluate meetingware. In: Proceedings of the 9th international workshop on groupware (CRIWG), vol 9, pp 109–125
3. Fjermestad J, Hiltz S (1999) An assessment of group support systems experimental research: methodology and results. J Manage Inf Syst 15(3):7–149
4. Karat J (1997) User-centered software evaluation methodologies. In: Helander M, Landauer P, Prahbu P (eds) Handbook of human–computer interaction. Elsevier, Amsterdam
5. Reinig BA, Shin B (2002) The dynamic effects of group support systems on group meetings. J Manage Inf Syst 19(2):303–325
6. Nunamaker JF Jr, Briggs RO, Mittleman DD, Vogel DR, Balthazard PA (1997) Lessons from a dozen years of group support systems research: a discussion of lab and field findings. J MIS 13(3):163–207
7. Jessup LM, Valacich JS (1993) Future directions and challenges in evolution of GSS. In: Jessup LM, Valacich JS (eds) Group support systems: new perspectives. MacMillan, New York
8. Neale DJ, Carroll JM, Rosson JM (2004) Evaluating computer-supported cooperative work: models and frameworks. Proc CSCW 112–121
9. Huis in 't Veld MAA (2007) E-Magine: an evaluation method to assess groups using ICT. Dissertation, Delft (in press)
10. Brodbeck FC (1996) Criteria for the study of work group functioning. In: West MA (ed) Handbook of work group psychology. Wiley, Chichester
11. McGrath JE, Hollingshead A (1994) Interacting with technology: ideas, evidence, issues and an agenda. Sage Publications, Thousand Oaks
12. West MA, Borrill CS, Unsworth KL (1998) Team effectiveness in organizations. In: Cooper CL, Robertson IT (eds) International review of industrial and organizational psychology, vol 13. Wiley, New York
13. Smith-Jentsch KA, Johnston JH, Paynes SC (1998) Measuring team-related expertise in complex environments. In: Cannon-Bowers JA, Salas E (eds) Making decisions under stress: implications for individual and team training. APA, Washington
14. Post WM, Cremers AHM, Blanson Henkemans OA (2004) A research environment for meeting behavior. In: Nijholt A, Nishida T (eds) Proceedings of the 3rd workshop on social intelligence design, University of Twente, Enschede
15. Guzzo RA, Salas E (1995) Team effectiveness and decision making in organizations. Jossey-Bass Publishers, San Francisco
16. Shaw ME (1971) Group dynamics: the psychology of small group behavior. McGraw-Hill, New York
17. Stewart GL, Manz CC, Sims HP (1999) Teamwork and group dynamics. Wiley, New York
18. Steiner ID (1972) Group process and productivity. Academic, New York

19. Schreiber G, Akkermans H, Anjewierden A, de Hoog R, Shadbolt N, van de Velde W, Wielinga B (2000) Knowledge engineering and management. MIT Press, Massachusetts
20. Short J, Williams E, Christie B (1976) The social psychology of telecommunication. Wiley, London
21. Swigger K, Brazile R, Peng X, Harrington B (2004) Computer-supported collaboration and the effects of culture. In: Darses F, Dieng R, Simone C, Zacklad M (eds) Supplement to the proceedings of cooperative systems design
22. Weick KE, Sutcliffe KM (2001) Managing the unexpected: assuring high performance in an age of complexity. Jossey-Bass (Wiley), San Francisco
23. Alblas G (1992) Groepsprestaties (Group performance). In: Meertens RW, von Grumbkow J (eds) Sociale psychologie (Social psychology). Wolters-Noordhoff, Groningen
24. Gaillard AWK (2003) Stress, productiviteit en gezondheid (Stress, productivity and health). Uitgeverij Nieuwerzijds, Amsterdam
25. Robbins SP (2001) Organizational behavior. Prentice-Hall, New Jersey
26. Jackson SE (1996) Consequences of diversity in multidisciplinary teams. In: West MA (ed) Handbook of work group psychology. Wiley, Chichester
27. Zijlstra F, Meijman T (1989) Het meten van mentale inspanning met behulp van een subjectieve methode (Measurement of mental effort with a subjective method). In: Meijman T (eds) Mentale belasting en werkstress: een arbeidspsychologische benadering (Mental workload and work stress: a work psychological approach). Van Gorcum, Assen
28. De Waard D (1996) The measurement of drivers' mental workload. Dissertation, University of Groningen, Traffic Research Center
29. van den Boogaard SAA (2005) Meeting behavior: exploring factors of success. Master thesis, University of Leiden
30. Post WM, Blanson Henkemans OA, van Verseveld OH (2006) Scenario definition. Report, TNO, Soesterberg
31. Laeser M, Moskal BM, Knecht R, Lasich D (2003) Engineering design: examining the impact of gender and the team's gender composition. J Eng Educ 92(1):49–56
32. Levine JM, Moreland RL (1990) Progress in small group research. Ann Rev Psychol 41:585–634
33. Fenwick GD, Neal DJ (2001) Effect of gender composition on group performance. Gender Work Organ 8(2):205–225
34. Punte PAJ, Hamel R (2001) Cognitieve beschrijving van het ergonomisch ontwerpen (Cognitive description of the ergonomic design process). Report, TNO, Soesterberg
35. Nieva VF, Fleishman EA, Reick A (1978) Team dimensions: their identity, their measurement, and their relationships. Response Analysis Corporation, Washington
36. Guzzo RA (1996) Fundamental considerations about work groups. In: West MA (ed) Handbook of work group psychology. Wiley, Chichester